Sentiment Classification in Social Media Data by Combining Triplet Belief Functions

Sentiment Classification in Social Media Data by Combining Triplet Belief Functions

Yaxin Bi

School of Computing
University of Ulster at Jordanstown
Co Antrim, BT37 0QB, UK

E-mail: y.bi@ulster.ac.uk
Telephone: +442890366582

Abstract

Sentiment analysis is an emerging technique that caters for semantic orientation and opinion mining. It is increasingly used to analyse online reviews and posts for identifying people's opinions and attitudes to products and events in order to improve business performance of companies and aid to make better organising strategies of events. This paper presents an innovative approach to combining the outputs of sentiment classifiers under the framework of belief functions. It consists of the formulation of sentiment classifier outputs in the triplet evidence structure and the development of general formulas for combining triplet functions derived from sentiment classification results via three evidential combination rules along with comparative analyses. The empirical studies have been conducted on examining the effectiveness of our method for sentiment classification individually and in combination, and the results demonstrate that the best combined classifiers by our method outperforms the best individual classifiers over five review datasets.

1 Introduction

In this big data era, social media, like forums, microblogging and social networking, is now pervasive in so many areas of people's social and economic lives. It provides the immense volumes of continuous real time data that collate timely opinions and views towards political events, feedback about products and quality of services, they deeply influence people's attitudes and beliefs (Feldman, 2013; Rodríguez-Vidal et al., 2019). Such views and opinions are thus a key source of helping politicians make election campaign strategies and aiding companies to monitor and predict customers' business behaviors, they play a fundamental role in the decision-making process for both individuals and organizations. However as a popular event and product often receive posts or reviews in the millions and in various forms of expressions, opinions conceived in reviews can be difficult to precisely be explained and identified. When used in analysing opinions and attitudes, such as the automated interpretation of reviews, sentiment classification will be imperative and extremely helpful. Sentiment classification is the core of sentiment analysis, it mainly concerned with making sense of reviews and determining positive, negative or neutral sentiment towards subjects. It also involves handling uncertainty entailed in determining polarity boundaries and identifying opinions contained in evolution of reviews.
Sentiment classification shares many aspects with text categorization in common. Both are typically developed on the basis of supervised machine learning approaches, i.e. using labeled text corpora to train classification models and then applying them to assign new text documents with the pre-defined topic/class labels. Text categorization is mainly focused on aspect/topic-based multiple class classification tasks, whereas sentiment classification is primarily interested in assigning opinionated texts into the categories of positive, negative or neutral against aspects/topics. In fact, the neutral sentiment does not indicate any preference or attitudes for a subject and less useful for any decision making tasks, hence it is not what people or companies are interested in. The inclusion of a neutral category in classification also increases a significant amount of costs in training classification models and makes it even challenging to construct opinionated text corpora labelled with the three categories from the proliferation of social media big data (Varathan et al., 2017).
Building on our previous studies (Bi, 2014; Bi et al., 2014; Burns et al., 2011, 2019), the aim of this work is to develop a dichotomous evidential ensemble approach for sentiment classification. We formulate positive and negative sentiment polarity as mutually exclusive propositions and sentiment classification results with the triplet evidential structure (Bi, et al. 2008), and then use a set comprising of positive and negative propositions to represent the neutral category in terms of *ignorance*. As such, the neutral sentiment is modelled with uncertainty entailed in the boundary of sentiment polarity and the degree of uncertainty is incorporated into the process of sentiment classification. This approach effectively minimizes the negative effect of the neutral sentiment through combining multi-classification models and performs classification tasks with an evidential reasoning mechanism. Our approach also saves the costs of constructing training corpora and building classification models. Moreover to effectively combine multiple triplet functions derived from different classification models together to infer sentiment polarity, we further develop the new computational formulas for the dichotomous situation based on Dempster's rule of combination (Shafer, 1976), the

Transferable Belief Model(TBM) conjunctive rule (Smets and Kennes, 1994) and the Cautious conjunctive combination rule (Denoeux, 2008).
This work contributes a systematic study on the development of formulation of sentiment classification and calculation formulas in conjunction with empirical analyses for sentiment classification under the theory of belief functions, which has not been appropriately addressed in either the area of sentiment analysis or the applications of ensemble machine learning. The existing studies on sentiment classification so far are in various forms, they used different versions of datasets with different experimental settings and the evaluation results are normally not directly comparable (Injadat et al., 2016; Ravi and Ravi, 2015; Hemmatian and Sohrabi, 2019; Wadawadagi and Pagi, 2020). In respect of the evaluation, we select five popular datasets and nine machine learning algorithms, and conduct substantial experiments in accordance with the well established convention for text categorization and supervised learning tasks, i.e. use of the bag of words (BoW) model to represent text reviews, a ten-fold cross validation along with F-score to measure the performance of the proposed method, in an effort to compare our results with the previous ones to the maximum extent possible.
The experiments comprise two groups of eight and nine learning algorithms conjointly with the three combination rules over the five popular review datasets, involving more than 11,000 classifiers training and aggregations of triplet belief functions. In the meanwhile, the experiments also include comparative studies against the baseline method of the Weighted Majority Voting and the ensemble method of AdaBoosting, as well as a state of the art deep learning algorithm of the Transformer Neural Network (TNN) (Nandan, 2020), these also involve more than 2,500 additional classifiers training, aggregations of triplet belief functions and comparative analysis over these data sets. The experimental results demonstrate the competitive performance of the proposed method.
The rest of the paper is structured as follows. In the first place, we briefly present related work in supervised learning approaches for sentiment classification and the important ideas about ensemble learning-based opinion mining in Section of Related work. After that Section of belief functions provides details of the formulation of sentiment polarity classification and computational formulas for combining the classifier outputs. Section of Evaluation presents an experimental setup and evaluation results showing the effectiveness of the proposed approach. In Section of Discussion we discuss experimental results and compare them against some existing work. Finally we present a summary of this study.

Related work

The relentless growth of the amount of information in social media has gained the increasing interest of research community to address sentiment analysis by two main approaches. The first approach is based on the use of machine learning techniques, involves building classifiers from labelled instances of texts or sentences (Pang et al., 2002), essentially a supervised classification task. The other uses lexical resources such as lexicons (also known as the unsupervised or semantic orientation approach) (Turney, 2002). This study focuses on the first approach and provides a concise summary of state of the art supervised machine learning methods and techniques used for sentiment classification and identifying major relevant issues that remain to be addressed.
The major applications of supervised machine learning to text categorization dates back to the early 90s. Sebastiani (2002) provided a comprehensive review on the principle methods and related aspects of supervised machine learning used for text categorization. Sentiment classification shares many common aspects with text categorization, somehow it can be seen as a special type of text categorization. The task of text categoriztion starts with text representation, where text documents are pre-processed and the words occurring in the documents are collected to form a vector space in accordance with the bag-of-words model. Subsequently assigning text documents to predefined categories is mainly based on the closeness between keyword vectors, where the importance of keywords in the documents is often measured by weighting schemes such as *word frequency* or*term frequency inverse document frequency* (*tfi-df*) (Apté et al., 1994; Lewis, 1992), the most common words and high frequent words are not included for classification.
By contrast, sentiment analysis has considered more linguistic features to capture meaningful units in

text documents. The n-gram is a popular model used in sentiment classification, it segments sentences into a set of string tokens with equal size and then takes the sequence of tokens into account in text representation, particularly when n=2 and 3, a n-gram is called bigram and trigram, respectively. The part of speech (POS) is widely used to tag words with grammatical properties that put words into different categories such as adjective, adverbs, verbs and noun. Those representations have been employed in many studies for sentiment classification (Pang and Lee, 2004; Pang et al., 2002; Xia et al. 2011). While sentiment assignment is primarily through measuring the closeness between the words occurring in documents or sentences. Those words are normally attributed with the linguistic properties of explicit positive or negative, such as 'good', 'bad', 'excellent', etc. they denote people's opinions and attitude towards products, services or events. In fact the frequency of words is less important in determining the polarity of documents, and some of words might be not used in the case of text categorization.

The incorporation of these fine-grained features into text representations aims at highlighting the importance of some words in a document in terms of occurrence along with the order in which the words occur in the document and the syntactic role they play. However a number of experiments conducted in text categorization and information retrieval revealed that more sophisticated feature representations could not yield significantly better performance (Apté et al., 1994; Lewis, 1992). In particular, some authors used individual words alongside phrases to representing text documents as indexing terms in a vector space (Schütze et al., 1995), but the experimental results were not uniformly encouraged. Likewise the empirical studies from sentiment classification also demonstrated similar results that the sophisticated and compound feature representations could not improve classification performance (Lane et al., 2012; Wang et al., 2014; Blair, 2020), but recent studies still have not ruled out the possibility that sophisticated and compound features are as equally useful as unigram presence or superior (Vilares, et al., 2015).

Yogarajan, et al. (2020) conducted a more comprehensive study on applications of word embedding using continuous bag of words (CBOW) methods in a healthcare context. The authors used two large datasets of MIMIC III with 4Getabye (GB) and TREC 2017 with 24GB to train their word embeddings with a range of parameters, i.e. window size 5, ten negative samples per positive sample and a various number of dimensions. Their experiments demonstrated that both an increase in embedding dimensionality and an increase in the volume of health-related training data could improve classification accuracy. As opposed to this finding a comparative study with the conventional BoW model showed that in many cases the BoW model is superior to the word embeddings, particularly in applications of sentiment analysis (Blair et al., 2020; Shuang, et al. 2020).

The work by Pang, et al. (2002, 2004) provided influential studies on applying machine learning algorithms to automatically classify online reviews with positive and negative polarity. They formulated sentiment classification problems based on conventional text categorisation tasks, i.e. representing text reviews and applying three well-known supervised learning algorithms of Naive Bayes (NB), Maximum Entropy (ME) and Support Vector Machines (SVM) to analyse the Internet Movie Database. They experimented sentiment classification with 8 feature representation models over a corpus of 752 negative and 1301 positive reviews, the results demonstrated that SVM yielded the highest accuracy of 82.9% with the unigram model. The authors also asserted that the traditional topic-based classification often identifies topics based on keywords alone, but sentiment requires more *understanding* of text to separate topics. This motivated a large number of follow up studies on incorporating linguistic properties into feature representation for sentiment classification.

Wang et al. (2014) carried out a comparative study on the ensemble learning methods of Bagging, Boosting and Random Subspace based on five learning algorithms, namely NB, ME, Decision Tree (DT), k-Nearest Neighbor (kNN), and SVM for sentiment classification, putting them together 20 combinations in total. The authors used a unigram feature model to represent text reviews and experimented on ten different datasets that were used in the study (Whitehead and Yaeger, 2009), their results demonstrated that the three ensemble methods achieved better accuracies compared with those base classifiers. Among twenty learning methods, Random Subspace SVM had the best accuracy. They suggested the performance of the ensemble methods could be further validated on large datasets.

Xia et al. (2011) provided a more comprehensive ensemble-based approach for combining different representations of features and classifiers. The authors conducted three investigations into representing

text features based on POS and word-relation (i.e. unigrams, bigrams and linguistic characteristics), three well-known text classification algorithms, namely NB, ME and SVM, as well as a set of classifier ensembles constructed by the three combination rules: fixed combination, weighted combination and meta-classifier combination. The authors particularly addressed the importance of syntactic relations by combining the word-relation features and a fast feature selection method. They evaluated the aspects of the feature representation and combination rules on five document-level datasets that were used by Pang and Lee (2004) and the data sets derived from Amazon that were used in the studies (Jo and Oh, 2011; Pang et al., 2002), the evaluation results showed that the performance of the classifier ensembles on the same feature set was robustly better than any individual classifiers and the weighted combination rule significantly outperformed the fixed combination and the meta-learning rules.

The work by Nguyen et al. (2018) developed an ensemble system for heterogeneous base classifiers with a fuzzy IF-THEN rule inference engine as the combiner to capture the uncertainty reflected in the outputs of the base classifiers. The fundamental idea of the proposed method was to use a set of fuzzy rules to operate on the outputs of base classifiers to make the final classification decision. The evaluation results on thirty datasets demonstrated the proposed method outperformed nine ensemble methods, including AdaBoost, Decision Template, Decision Tree on meta-data, and so forth. However the claim of capturing uncertainty from the agreements and disagreements between the different base classifiers is not convincing, it is unclear how the classifier outputs could be treated as meta-data that is viewed as a transformation from feature domain to posterior probability domain with all value in the range of 0 and 1.

In addition to the relevant studies above, there are a body of recent surveys on sentiment analysis. Ravi et al. (2015) reviewed the 161 articles published in the period of 2002-2015, they organized the articles into the categories of machine learning, natural language processing techniques and applications of sentiment analysis. The authors summarized public available datasets, various types of syntactic, semantic, and statistical feature representations in conjunction with machine learning algorithms, and applied approaches for sentiment analysis. They also identified 20 intelligent techniques in the machine learning category including SVM, NB, DT, etc. and ranked SVM as the top performance algorithm.

Varathan et al. (2017) conducted a survey on applications of machine learning, rule mining and natural language processing techniques on opinion mining. The survey included 38 articles reported in the period of 2006–2015. The authors systematically compared the applications of those techniques of detecting sentiment at the element levels of sentence, entity, relation and feature, which cover the two perspectives of techniques and comparative opinion elements. They also presented preprocessing tools and data sets used along with potential development of the area, these results provided insights into future researches in the field of sentiment analysis.

The more recent advancement of deep learning approaches has inspired a number of applications for sentiment analysis. Socher et al. (2013) proposed a Recursive Deep Models for capturing semantic compositional effects over a sentiment treebank. The underlying part of that model is a new called Recursive Neural Tensor Network (RNTN), where it parses an input document into a set of phrases that are represented by word vectors and a parse tree, and then calculates the vectors for higher nodes along the tree using the tensor-based composition function. Their experimental results demonstrated a considerable improvement in single sentence positive/negative classification from 80% up to 85.4% over the corpus of movie review used by Pang and Lee (2005).

Furthermore, Wadawadagi, et al. (2020) have conducted a comparative study and performance assessment on sentiment analysis with deep neural networks. The authors reviewed more than 130 recent articles and concluded that although deep learning algorithms have shown significant outcomes in sentiment analysis, yet these algorithms have some drawback, for instance, Convolutional Neural Network (CNN) requires tuning of initial hyperparameters as a starting point, which was also criticized by Stojanovski et al. (2015). Hence the performance of deep learning models depend upon the values of the hyperparameters, but determining a set of optimal hyperparameter values is a challenging task. When processing very long sequences, those models often suffer from the vanishing gradient and exploding gradients problems, therefore making the deep learning approaches unfavorable for text analysis tasks.

Summarising the relevant aforementioned studies, we have identified three issues that are compelling to be addressed in this area as follows:

- The literature has showed that no single sentiment method always achieved the best performance for different datasets. There are also few studies of combining classifier results for sentiment classification by explicitly applying well-established fusion techniques. Thus more extensive experimental work is needed, particularly applying evidential reasoning approaches to the combination of classifiers for sentiment classification.
- Determining sentiment classification requires to handle uncertainty entailed in the boundaries among positive, negative and neutral categories. Currently there is lack of effective ways to cope with such uncertainty in determining the boundaries of sentiment categories particularly considering sentiment classification as a binary classification task.
- The methodology for evaluating machine learning and data mining algorithms has been well established in the literature, such as k-fold cross validation, F-measure, statistical reliability of estimated performance, and kappa κ statistic coefficient of inter-rater agreement. However quite a few articles demonstrate appropriate utilization of the methodology, thereby making comparative analyses more difficult to be conducted.

In this study, we primarily focus on the above issues, applying the conventional supervised learning algorithms to sentiment classification individually and in combination, especially in the context of an evidential reasoning framework.

Basics of belief functions

This section briefly introduces some essential concepts and formalism of the Dempster-Shafer (DS) theory of evidence, which are used in this study. Given a frame of discernment, denoted by Θ that consists of mutually exclusive propositions, for any subset $A \subseteq \Theta$, the DS theory uses a numeric value within the range of [0, 1] to represent the support degree for A, denoted by $m(A)$, which is called a **Basic Belief Assignment** (BBA) (Shafer, 1976) . It has two conditions:

$$(1)\ m(\emptyset) = 0; \quad (2)\ \sum_{A \subseteq \Theta} m(A) = 1$$

where A is called a focal element or focus if $m(A) > 0$, and it is called a singleton if A contains only one element with $m(A) > 0$.
Notice that $m(A)$ represents the measurement of a support degree that one commits exactly to the subset A, rather than the total support committed to A. To obtain the measurements of the total support committed to A and of other situations, the DS provides three other evidential functions in terms of **belief function** (bel), **plausibility function** (pls), and **commonality function** (q). Particularly the computation of a commonality function for subset $A \subseteq \Theta$ by a mass function is given as follows:

$$q(A) = \sum_{B \supseteq A} m(B) \tag{1}$$

In practice, evidence sources may not be entirely reliable. Support degrees derived from such sources need to be discounted to truly reflect the reliability of the sources. The DS theory provides a discount mechanism by the following formula.

$$m^r(A) = \begin{cases} (1-r)m(A), & \text{if } A \subset \Theta \\ r + (1-r)m(\Theta), & \text{if } A = \Theta \end{cases} \tag{2}$$

where r represents a **discounting rate** within the range of $[0, 1]$. When an evidence source is completely reliable, r takes on the value of 0. As opposed to this, if the source is completely unreliable, r takes the value of 1.
Definition 1 Let m_1 and m_2 be two mass functions on the frame of discernment Θ, and for any subset $A \subseteq \Theta$, the *orthogonal sum* of two mass functions on A, denoted by $\oplus$, is defined as follows:

$$m_1 \oplus m_2(A) = \frac{\sum_{X \cap Y = A} m_1(X) m_2(Y)}{1 - \sum_{X \cap Y = \emptyset} m_1(X) m_2(Y)} \tag{3}$$

where $X, Y \subseteq \Theta$ and $1 - \sum_{X \cap Y = \emptyset} m_1(X) m_2(Y)$ is a normalization factor, denoted by K. The orthogonal sum is also called Dempster's rule of combination, it allows two mass functions to be aggregated to a third mass function. The above definitions are based on the notations given by Shafer (1976). In the framework of the **Transferable Belief Model** (TBM) (Smets and Kennes, 1994), Smets lifts the restriction on $m(\emptyset) = 0$ and removes the normalization operator in Equation (4), resulting in a TBM conjunctive combination rule, denoted by $\textcircled{\cap}$. Both combination rules have been employed in the Cautious conjunctive combination rule developed by Denoeux (2008).

Definition 2 Let m_1 and m_2 be two non dogmatic[1] mass functions on the frame of discernment Θ. The Cautious conjunctive combination of m_1 and m_2, denoted by $m_{1\textcircled{\wedge}2} = m_1 \textcircled{\wedge} m_2$, is defined on the basis of a weight function below:

$$w_{1\textcircled{\wedge}2}(A) = w_1(A) \wedge w_2(A), \forall A \subset \Theta \tag{4}$$

and then

$$m_1 \textcircled{\wedge} m_2 = \textcircled{\cap}_{A \subset \Theta} A^{w_1(A) \wedge w_2(A)}. \tag{5}$$

where $A^{w(A)}$ denotes a **simple mass function**[2] m such that $m(A) = 1 - w(A)$ and $m(\Theta) = w(A)$, and the weight $w(A)$ can be obtained from the commonality values by the following formula:

$$w(A) = \prod_{B \supseteq A} q(B)^{(-1)^{|B|-|A|+1}} \tag{6}$$

if the conjunctive operation $\textcircled{\cap}$ in Equation (6) is replaced by the Dempster orthogonal sum $\bigoplus$, then the revised Equation (6) has to be divided by the normalization factor of $K = 1 - m_1 \textcircled{\wedge} m_2(\emptyset)$, resulting in a normalized version of the Cautious rule as follows:

$$m_1 \textcircled{\wedge} m_2 = \frac{\bigoplus_{\emptyset \neq A \subset \Theta} A^{w_1(A) \wedge w_2(A)}}{K}. \tag{7}$$

As indicated by Denoeux (2008) that $m_1 \textcircled{\wedge} m_2 = 1$ never holds as **the Cautious combination of two non dogmatic BBAs can never be dogmatic** (i.e. Θ *is not a focal set*).

Triplet and binary structures

Sentiment polarity classification is simply treated as a special case of topic-based categorization with positive and negative polarities. In such a case, a categorization algorithm is provided with a training data set made up of $D \times C = \{\langle d_1, c_1\rangle, \cdots, \langle d_{|D|}, c_q\rangle\} (1 \leq q \leq |C|)$ for deriving a classifier denoted by φ. A text review $d_i \in D$ is characterized by the vector of words of $(\omega_{i_1}, \cdots, \omega_{i_n})$ where ω_{i_j} is a score of keyword, C is composed of positive polarity (c) and negative polarity ($\tilde{c}$), i.e. $C = \{c, \tilde{c}\}$, and a classifier output on a new review d is denoted by $\varphi(d) = C \times [0, 1]$. Given classifier output $\varphi(d)$, we can formulate it as a piece of evidence in the form of triplet below.

Definition 3 Let $\Theta = \{x_1, x_2, ..., x_n\}$ be a frame of discernment and $\varphi(d) = \{m(\{x_1\}), \cdots, m(\{x_n\})\}$ be mass probabilities derived from classifier outputs, where $n \geq 2$. The expression of $A = \langle A_1, A_2, A_3\rangle$ is defined as a *triplet*, where $A_1, A_2 \subset \Theta$ are singletons, A_3 is the whole set Θ and they satisfy

$$m(A_1) + m(A_2) + m(A_3) = 1$$

[1] meaning that the frame of discernment Θ is not a focal set.

[2] comprising only one focal element and the frame of discernment Θ.

When $n = 2$, the frame of discernment Θ is composed of only two focal elements, denoted by $\{x\}$ and $\{\tilde{x}\}$, i.e. $\Theta = \{x, \tilde{x}\}$, such that a triplet mass function defined on Θ also satisfies the following condition:

$$m(\{x\}) + m(\{\tilde{x}\}) + m(\Theta) = 1$$

and the new support degrees for $\{x\}$, $\{\tilde{x}\}$ and Θ can be obtained by using normalization and the discounting Equation (3).

Combining triplet mass functions via the evidential combination rules

Following the above section, this section formulates the combinations of two mass functions committed to binary focal elements by using Dempster's rule, Smets's rule and the Cautious conjunctive rule, which can be tailored to compute the combinations of any number of triplet functions.

Definition 4 Let $\Theta = \{x, \tilde{x}\}$ be a frame of discernment, and m_1 and m_2 be triplet mass functions with the following condition:

$$m_1(\{x\}) + m_1(\{\tilde{x}\}) + m_1(\Theta) = 1$$
$$m_2(\{x\}) + m_2(\{\tilde{x}\}) + m_2(\Theta) = 1$$

combining m_1 and m_2 by using Smets's rule, we have the combined result below:

$$m(\{x\}) = m_1(\{x\})m_2(\{x\}) + m_1(\{x\})m_2(\Theta) + m_1(\Theta)m_2(\{x\}) \quad (8)$$

Equation (9) can be then rewritten as Equation (10) with a **normalization factor** as combined by Dempster's rule below

$$m(\{x\}) = [(m_1(\{x\}) + m_1(\Theta)) \times (m_2(\{x\}) + m_2(\Theta)) - m_1(\Theta)m_2(\Theta)]/K \quad (9)$$

where $K = 1 - m_1(\{x\})m_2(\{\tilde{x}\}) - m_1(\{\tilde{x}\})m_2(\{x\})$
Likewise we can have new mass functions for $\{\tilde{x}\}$ and Θ below:

$$m(\{\tilde{x}\}) = [(m_1(\{\tilde{x}\}) + m_1(\Theta)) \times (m_2(\{\tilde{x}\}) + m_2(\Theta)) - m_1(\Theta)m_2(\Theta)]/K \quad (10)$$

$$m(\Theta) = m_1(\Theta)m_2(\Theta)/K \quad (11)$$

In respect of combining two triplet mass functions m_1 and m_2 by using the Cautious conjunctive rule, it involves three steps (Denoeux, 2008). The first step computes the commonality functions q_1 and q_2 from m_1 and m_2 by Equation (2); the second computes the weight functions w_1 and w_2 using Equation (7) and then generate (inverse) simple mass functions in the form of $A^{w_1(A) \wedge w_2(A)}$, for all $A \subset \Omega$ such that $w_1 \wedge w_2(A) \neq 1$; and finally computes $m_1 \textcircled{\wedge} 2 = m_1 \textcircled{\wedge} m_2$ on these simple mass functions by using Smets's rule. Following the first two steps, we finally have,

$$\{x\}^{w_1(\{x\}) \wedge w_2(\{x\})} = \{x\}^{min(\{\frac{m_1(\Theta)}{m_1(\{x\})+m_1(\Theta)}, \frac{m_2(\Theta)}{m_2(\{x\})+m_2(\Theta)}\})=w}$$

$$\{\tilde{x}\}^{w_1(\{\tilde{x}\}) \wedge w_2(\{\tilde{x}\})} = \{\tilde{x}\}^{min(\{\frac{m_1(\Theta)}{m_1(\{\tilde{x}\})+m_1(\Theta)}, \frac{m_2(\Theta)}{m_2(\{\tilde{x}\})+m_2(\Theta)}\})=\tilde{w}}$$

with w and $\tilde{w}$, we can generate two (inverse) simple mass functions as follows,

$$m^1_{12}(\{x\}) = 1 - w; m^1_{12}(\Theta) = w.$$

$$m_{12}^2(\{\tilde{x}\}) = 1 - \tilde{w}; m_{12}^2(\Theta) = \tilde{w}.$$

From the last step above, we have

$$m(\{x\}) = m_{12}^1(\{x\}) m_{12}^2(\Theta). \tag{12}$$

$$m(\{\tilde{x}\}) = m_{12}^2(\{\tilde{x}\}) m_{12}^1(\Theta). \tag{13}$$

$$m(\Theta) = m_{12}^1(\Theta) m_{12}^2(\Theta). \tag{14}$$

Example: Let us illustrate an application of the concepts and formulas above in combining two sentiment classifiers. Suppose we have two classifiers of NB and SMO, given an online review, these classifiers produce two outputs over the positive and negative categories. With the triplet formulation, we denote the positive category by x and negative by $\tilde{x}$, m_1 and m_2 represent the support degrees over x, $\tilde{x}$ and $\Theta = \{x, \tilde{x}\}$, which are derived from the two outputs by Equation (3). Given these settings, we now assume two triplet belief functions as follows:

$$m_1(\{x\}) = 0.70048, m_1(\{\tilde{x}\}) = 0.21293, m_1(\Theta)) = 0.08659$$
$$m_2(\{x\}) = 0.14726, m_2(\{\tilde{x}\}) = 0.68456, m_2(\Theta)) = 0.16818$$

Combining these two triplet functions by Equations (10) , (11) and (12), we obtain the support degrees over the positive and negative categories below:

$$m_{12}(\{x\}) = 0.478, m_{12}(\{\tilde{x}\}) = 0.492, m_{12}(\Theta) = 0.0298$$

where $m_{12}(\Theta)$ represents the neutral category. In accordance with the maximum of the support degrees, this review is assigned to the negative category.
Given multiple triplet mass functions of $m_1, \ldots, m_L$ that are committed to binary focal elements $\{x\}$ and $\{\tilde{x}\}$, the general formulas of combining them by using Smets's rule can be obtained below (Bi, 2008):

$$m(\{x\}) = \prod_{i=1}^{L}[m_i(\{x\}) + m_i(\Theta)] - \prod_{i=1}^{L} m_i(\Theta) \tag{15}$$

$$m(\{\tilde{x}\}) = \prod_{i=1}^{L}[m_i(\{\tilde{x}\}) + m_i(\Theta)] - \prod_{i=1}^{L} m_i(\Theta) \tag{16}$$

$$m(\Theta) = \prod_{i=1}^{L} m_i(\Theta) \tag{17}$$

Given the general formula above, the **normalization factor** for Dempster's rule can be derived as follows

$$K = \prod_{i=1}^{L}[1 - m_i(\{x\})] + \prod_{i=1}^{L}[1 - m_i(\{\tilde{x}\})] - \prod_{i=1}^{L} m_i(\Theta) \tag{18}$$

When dividing Equations (16) - (18) by the factor K, we obtain the general Dempster's rule of combination.
Moreover to combine $m_1, \ldots, m_L$ by the Cautious rule, we arrange these L functions into a group of pairs as the result of combining each pair is also a triplet mass function, thus we can combine each pair using Equations (13) - (15) repeatedly as follows:

$$[\ldots[[[m_1 \circledwedge m_2] \circledwedge m_3] \circledwedge m_4] \circledwedge \ldots \circledwedge m_L] \tag{19}$$

$$c = argmax_{x,\tilde{x} \in \Theta}\{m(\{x\}), m(\{\tilde{x}\})\} \tag{20}$$

By Equations (13)- (20), we can combine all multiple triplet mass functions defined on the binary frame of discernment to construct classifier ensembles for sentiment classification. Final classification decision will be made on the basis of Equation (21).

Table 1: Summary of datasets used for evaluation

Dataset	#Reviews	#positive	#negative	#words per review
Amazon	22354	17422	4932	182
LMR	25000	12500	12500	233
Rotten	9606	4898	4708	21
Trip	30212	22554	7658	152
Yelp	20708	17457	3252	158

Experimental Evaluation

Experiment settings

To evaluate the effectiveness of sentiment ensemble classifiers constructed by Dempster's rule of combination (Dempster's rule), the TBM conjunctive combination rule (Smets's rule) and the Cautious conjunctive rule (Cautious rule) against the baselines of Weighted Majority Voting (Wvoting) and the AdaBoosting method, we conducted the experiments with nine machine learning algorithms for generating base classifiers over five popular benchmark datasets. These algorithms include 1: NaiveBayes (NB), 2: Support Vector Machine (SMO), 3: IBk, 4: KStar, 5: DecisionStump (DSt), 6: J48, 7: RandomForest (RF), 8: DecisionTable (DT) and 9: JRip, which are directly taken from the Waikato Environment for Knowledge Analysis (Weka) version 3.6 (Witten et al., 2011), and the parameters used in these algorithms for the experiments were equally set at the default settings for a fair comparison.

Five popular benchmark data sets were selected for evaluating our methods, including Amazon (Jo and Oh, 2011), Large Movie Review (IMDB, 2006), Rotten-IMDB (Pang and Lee, 2004), TripAdvisor (Tripadvisor, 2006) and Yelp (Jo and Oh, 2011). For the Amazon, LMR, TripAdvisor and Yelp datasets, each of the reviews encompasses short sentences and is labeled with a five star scale. In order for polarity classification, we convert the labelled data into three categories as done by Pang et al. (2002). Like the categorical judgment given by Zhitomirsky-Geffet et al. (2018), the reviews labelled with 1 and 2 stars are considered very negative and negative, whereas reviews with 4 and 5 stars are considered positive and very positive, and reviews with 3 stars are regarded as neutral (Pang et al., 2002). For this study, we confine our focus only on discriminating between positive and negative polarity sentiment. Therefore after pre-processing and removing neutral reviews, we collected 107880 reviews in total. Table 1 presents a summary of datasets used in the empirical study.

Although there are numbers of ways for feature representation, such as n-gram and incorporation of linguistic characteristics in feature representation, the BoW model has dominated the text categorisation related literature for past decades. For the aim of this study, we simply followed the empirical results from the work (Wang et al., 2014) and used information gains to select about 5000 keywords from these datasets, treating each review as a bag of words (unigrams) in the form a vector that comprises of presence/absence of all the keywords occurring in these datasets.

To construct classifier ensembles by the three evidential combination rules, the prerequisite is to formulate the outputs of classifiers as probabilities in the form of $\varphi(d) = \{c, \tilde{c}\} \times [0, 1]$, a triplet mass function. By using the evidential combination rules, all combinations of classifiers are exhaustively permuted. Specifically we first combine any two classifiers, denoted by 2C, and combine the resulting combination of two classifiers with a third classifier, denoted by 3C, and the result with a fourth classifier, denoted by 4C, until combine all classifiers.

However the SVM algorithm implemented in Weka only yields 0 and 1 binary outputs, which cannot be formulated as triplet mass functions to combine with the triplet mass functions derived from the other numeric classifiers. By the discount function Equation (3), we transform the binary outputs from the SVM into uniform triplet mass functions. Due to this exceptional treatment and examining the effect of the uniform functions, we divide our experiments into two groups, in which the first group of the experiments involves eight learning algorithms excluding SVM and the second group include all the nine algorithms.

To comply with the conventional experiment methodology of supervised learning methods and to avoid

Table 2: The accuracies (F-score) of best individual classifier (BI), best classifier ensembles constructed by Dempster's rule (DR), Smets's rule (SR) and the Cautious conjunctive rule (CR) on TripAdvisor (Trip), Amazon, Yelp, Large Movie Review and Rotten-IMDB (Rotten) in respect of eight classifiers.

Dataset	**BI**	**DR**	**SR**	**CR**
Amazon	77.74 (**1**)	83.63 (**1, 6, 7**)	83.75 (**1, 6, 7**)	81.13 (**1, 6**)
LMR	83.19 (**1**)	83.87 (**1, 7, 8, 9**)	83.78 (**1, 7, 8, 9**)	83.61 (**1, 7**)
Rotten	82.58 (**7**)	83.09 (**5, 7**)	83.09 (**5, 7**)	83.09 (**5, 7**)
Trip	87.03 (**1**)	89.39 (**1, 7, 8**)	89.48 (**1, 6, 7, 8**)	88.66 (**1, 6**)
Yelp	74.45 (**1**)	77.25 (**1, 7, 8**)	77.62 (**1, 6, 7**)	77.21 (**1, 6**)
Av	81.00%	83.45%	83.55%	82.74%

Table 3: The accuracies of best individual classifier (BI), best classifier ensembles constructed by Dempster's rule (DR), Smets's rule (SR), the Cautious conjunctive rule (CR) and the Weighted Voting (WV) on TripAdvisor (Trip), Amazon, Yelp, Large Movie Review and Rotten-IMDB (Rotten) in respect of nine classifiers.

Dataset	**BI**	**DR**	**SR**	**CR**	**WV**
Amazon	86.02 (**2**)	86.37 (**1, 2, 7, 8**)	86.38 (**1, 2, 6, 7**)	86.02 (**2, 8**)	86.17 (**1, 2, 6**)
LMR	83.85 (**2**)	85.49 (**1, 2, 6, 7**)	85.52 (**1, 2, 6, 7, 8**)	85.19 (**1, 2**)	85.23 (**1, 2, 6**)
Rotten	85.37 (**2**)	86.22 (**2, 4, 6, 7**)	86.04 (**2, 4, 6, 7**)	85.52 (**2, 7**)	85.75 (**2, 6, 7, 8**)
Trip	91.13 (**2**)	91.68 (**1, 2, 7, 8**)	91.71 (**1, 2, 5, 7, 8**)	91.13 (**2, 7**)	91.46 (**1, 2, 8**)
Yelp	81.71 (**2**)	82.02 (**1, 2, 8**)	81.84 (**1, 2, 8**)	81.71 (**2, 9**)	81.76 (**1, 2, 8**)
Av	85.61%	86.35%	86.30%	85.91%	86.07%

overfitting to some extent, the experiments were conducted using a ten-fold cross validation. The performance of classifiers in individual and combination was measured by the F-measure that is commonly used in text categorization area.

Experimental results

Table 2 summarises the accuracies of the best individual classifiers and the best classifier ensembles constructed by the three combination rules on the five datasets. Compared with the best individuals, the accuracies of the best ensembles increase by 2.45% for Dempster's rule, 2.55% for Smets's rule and 1.74% for the Cautious rule compared with the best individuals on average. These give a 2.45% increase averaged on the eight classifiers over the five datasets. In addition, alongside the accuracies of the best individual classifiers and classifier ensembles, the numbers in baskets indicate which classifier or ensemble performs best, which correspond to those numbers of the algorithms given in Subsection of Experimental Setting. In respect of the Amazon, for instance, (1) in the **Best individual** column means the classifier is generated by NB, (1, 6, 7) in the **DR** column means the classifier ensemble constructed by Dempster's rule is composed of three classifiers of NaiveBayes, RandomForest and DecisionTable.

Table 3 presents the accuracies of the best individual and best classifier ensembles over the five datasets in the case of nine classifiers. In comparison with the best individuals, Dempster's rule performs slightly better than Smets's rule and the Cautious rule. The best improvement of the best classifier ensembles is by 1.67%. The averaged accuracy of the best classifier ensembles over the five datasets increases by 0.74% for Dempster's rule, 0.69% for Smets's rule and 0.3% for the Cautious rule. Compared with the baseline of the Majority Voting, the averaged accuracy of the best classifier ensemble made by Dempster's rule is 0.28% better, by Smets's rule it is 0.23% better, but the Cautious rule does not outperform the Voting rule.

To analyse the performance of the best classifier ensembles, we present the performance of the classifier ensembles constructed by the three evidential rules in an increment order, examine their performance variations and then compare with the baseline rule of the Majority Voting.

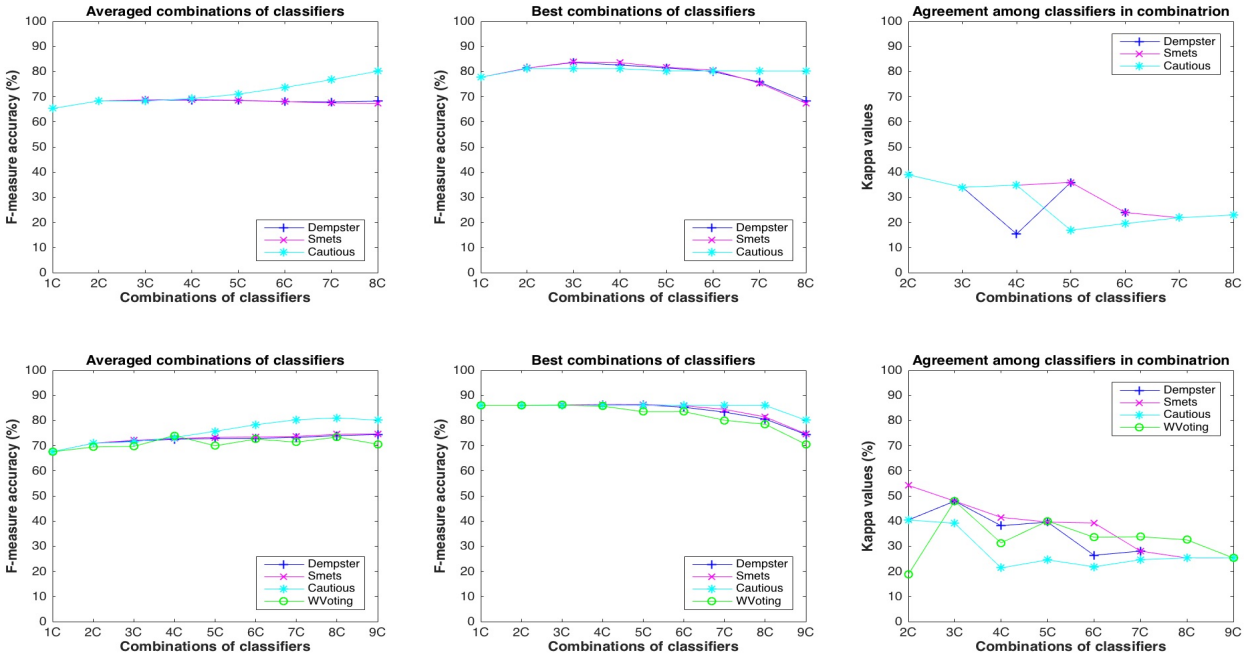

Figure 1: Accuracies of combined classifiers by four combination rules and respective Kappa values over the Amazon dataset

Experimental results of Amazon

Figure 1 presents the two groups of experimental results on the Amazon dataset. The upper row shows the first group of results in the case of eight classifiers, consisting of the performance of the averaged classifier ensembles, the best classifier ensembles and the κ statistics of the member classifiers making up the best ensembles. The lower row shows the second experiment results in the same setting as the upper row about the performance along with the performance of the baseline rule of of the majority voting, and κ statistics in the case of nine classifiers. In the first column of the diagram: averaged classifier ensembles, 1C on the x-axis represents the combinations of eight /nine individual classifiers and the respective averaged accuracy of the individual classifiers is presented on y-axises, 2C represents the classifier ensemble that is comprised of two classifiers by using the four combination rules and the y-axises represent the averaged accuracy of the corresponding classifier ensembles, 3C represents the ensemble made up of three classifiers by the combination rules, and so forth. These upper and lower diagrams illustrate that the averaged accuracies of different groups of the classifier ensembles gradually increase with adding more classifiers into the ensembles, particularly for those made up of some from the nine classifiers. It also can be observed that the averaged performance of the combined classifiers by Dempster's rule is very similar to that of the classifier ensembles constructed by Smets's rule and Majority Voting, the ascending trend of the classifier ensembles made by the Cautious rule appears to be more bullish than those of the classifier ensembles made by Dempster's, Smets's and Majority Voting rules. Compared with the averaged accuracy of eight base classifiers, the averaged accuracy of all classifier ensembles made by Dempster's rule is 3.02% better, by Smets's rule it is 2.91% better, and by the Cautious rules it is 7.2% better. In comparison with the averaged accuracy of nine base classifiers, the averaged accuracy of all classifier ensembles by Dempster's rule are 5.24% better, by Smets's rule it is 5.60% better, by the cautious rules it is 8.81% better, and by the Majority Voting rule it is 3.79% better. In the both average cases, the cautious rule performs best.
In the second column of the diagram: best classifier ensembles, 1C, 2C, ..., 8C/9C on the x-axis represent the best combinations of classifiers, which are made up of the best individual classifiers, two classifiers, three classifiers until eight/nine classifiers. Unlike the performance trends of the averaged

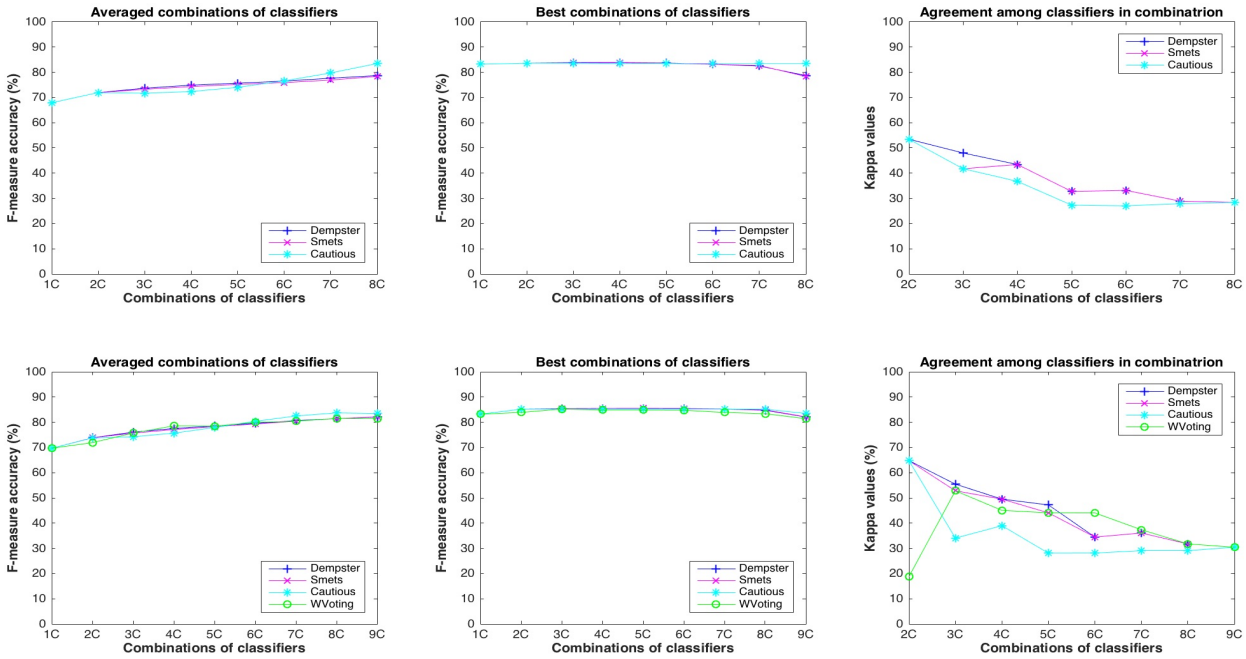

Figure 2: Accuracies of combined classifiers by four combination rules and respective Kappa values over the Large Movie Review dataset

classifier ensembles, the accuracies of the best classifier ensembles increase from the combination of two classifiers to five classifiers and then drop down when adding more classifiers into the classifier ensembles, whereas these made by the Cautious rule have almost flat performance. It can be found that the classifier ensembles made by Smets's rule perform the best and by Dempster's rule perform very similar in the both cases. The best accuracy drawn from all the classifier ensembles made by Smets's rule is 6.01% better than the best individual classifier, by Dempster's rule it is 5.9% better, and by the Cautious conjunctive rule it is 3.39% better in the case of eight classifiers. Compared with the best individual of nine classifiers, the best classifier ensemble does not outperform the best individual with a large margin, the accuracy of the best ensemble constructed by Smets's rule is 0.36% better, by Dempster's rule it is 0.35% better and for the Majority Voting it is 0.15% better.
The last column in Figure 1 illustrates the κ statistics, representing the extent of pairwise agreement among the classifiers in the best classifier ensembles, the strength of the agreement among these classifiers is generally fair at a 95% confidence interval, while some cases of the combinations of the nine classifiers have the moderate strength of the agreement. The fair agreement implies that when classifying the Amazon reviews in combination, the member classifiers tend towards the same classification decisions in a chance-corrected manner. In other words they agree one another in classifying the reviews to fair extent. It also can be seen that the general trend of the agreement is decreasing when combining more classifiers, this pattern is consistent with the intuition. In both case of eight and nine classifiers, the agreement among the classifiers making up the best classifier ensembles by Smets's rule is stronger than that of these classifiers that form the best ensembles by Dempster's rule, the Cautious rule and the Majority Voting, the extent of the agreement among the classifiers combined by Dempster's rule is similar to that of the classifiers made by the Majority Voting, but stronger than that of those made by the Cautious rule.

Experimental results of LMR

Figure 2 has the same structure and notations as illustrated in Figure 1, presenting the performance of the averaged classifier ensembles and best classifier ensembles in conjunction with the κ statistics on

the LMR dataset. Looking at the first column of the figure, the averaged classifier ensembles composed in the case of eight classifiers perform slightly different from those made up from the nine classifiers but with the same trend, that is when more classifiers being combined, the averaged performance of the classifier ensembles increases, by contrast the classifier ensembles constructed from the nine classifiers appear to increase monotonously. Meanwhile, the classifier ensembles constructed by the Cautious rule reach the best accuracy, performing better than those constructed by Dempster's rule and Smets' rule. Compared with the averaged performance of the eight classifiers, the averaged accuracy of the classifier ensembles made by Dempster's rule is 7.67% better, by Smets's rule it is 7.21% better and by the Cautious rule it is 7.78% better. In the case of nine classifiers, the averaged performance of the classifier ensembles made by Dempster's rule is 9.03% better than that of the nine classifiers, by Smets's rule it is 8.84% better, by the Cautious rule 9.27% better and by the Majority Voting 8.81% better.
The second column of the figure presents the best combinations of eight/nine classifiers by the three different combination rules. The construction of the best classifier ensembles made by the Cautious rule does not have clear effect on increasing the performance of the best ensembles in the both cases. The performance of the classifier ensembles made by the three combination rules appear to be compact and very similar before the combination of six classifiers (6C), which the performance of the best classifier ensembles appears to be flat, i.e. from one to the combination of two classifiers and drops after combining six classifiers by Dempster's and Smets's rules. In respect of the best ensembles composed from the nine classifiers, the performance increases from the best individual to the combination of three classifiers and the performance of the best ensembles drops after the combination of seven classifiers. The accuracy of the best classifier ensembles made from eight classifiers by Dempster's rule is 0.68% better than the best individual, by Semts's rule it is 0.59% better and by the Cautious rule the accuracy of the best ensembles is 0.42% better. In the case of nine classifiers, the accuracy of the best ensembles made by Dempster's rule is 1.64% better than the best individual, by Smets's rule it is 1.67% better, the Cautious rule is 1.34% better and by the Majority Voting it is 1.38% better, which outperform the case of eight classifiers.
The κ curves in Figure 2 delineate the extent of agreement among the member classifiers making up the best classifier ensembles in classifying the reviews. The degrees of the agreement among the classifiers decease when more classifiers are combined in the both cases of eight and nine classifiers with the exception of the Majority Voting rule. The small fluctuations occur over the combination of four classifiers by Smets's rule in the case of eight classifiers, whereas a similar phenomenon occurs in the combinations of two classifiers by the Majority Voting and three classifiers by the Cautious rule in the case of nine classifiers. Furthermore, the classifiers that make up the best classifier ensembles by Dempster's and Smets's rules have a similar agreement, which is stronger than that made by the Cautious rule and baseline rule of Majority Voting in the both cases. The insight here is that the decreasing of κ values when combining more classifiers is not only related to number of classifiers, but also the composition of the classifier ensembles.

Experimental results of Rotten-IMDB

Following the same structure as above, Figure 3 presents the performance of the classifier ensembles by the three combination rules along with the κ statistic curves over the Rotten-IMDB dataset. The averaged performance of the classifier ensembles made by Dempster's and Smets's rules shows a similar pattern in the both case of eight and nine classifiers, the accuracies increase throughout the combinations, and the ensembles made from nine classifiers perform better than those with eight classifiers. As opposed to this pattern, the averaged classifier ensembles made from eight and nine classifiers by the Cautious rule perform differently, the moderate increase is in contrast to a decrease when adding more classifiers to construct the classifier ensembles. The averaged performance of the classifier ensembles made from eight classifiers by Dempster's rule is 10.61% better than the averaged accuracy of the individual classifiers, by Smets's rule it is 9.09% better, and by the Cautious rule it is 2.55% better. Compared with the averaged accuracy of the nine classifiers, the ensembles made by Dempster's rule perform10.35% better, by Smets's rule they do 9.83% better, by the Cautious rule they perform 4.77% better and by the Majority Voting 4.86% better.
With respect to the best classifier ensembles in the second column, the performance of the best classifier

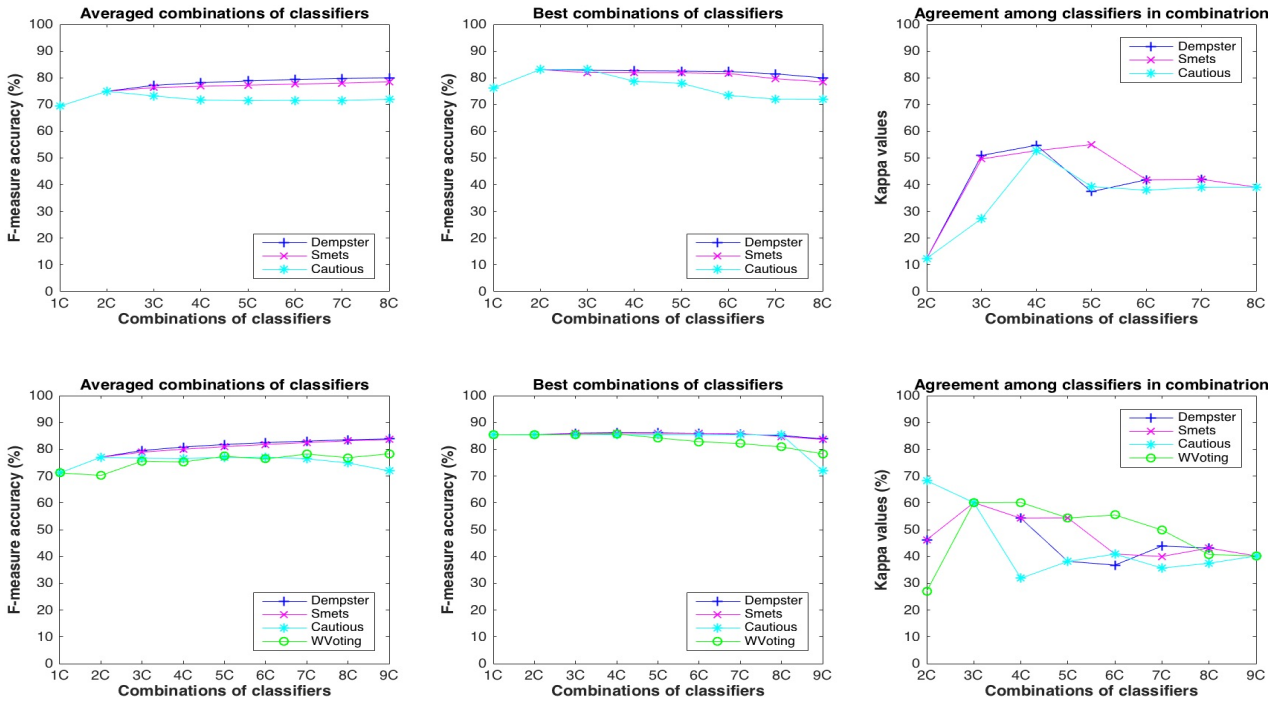

Figure 3: Accuracies of combined classifiers by four combination rules and respective Kappa values over the Rotten-IMDB dataset

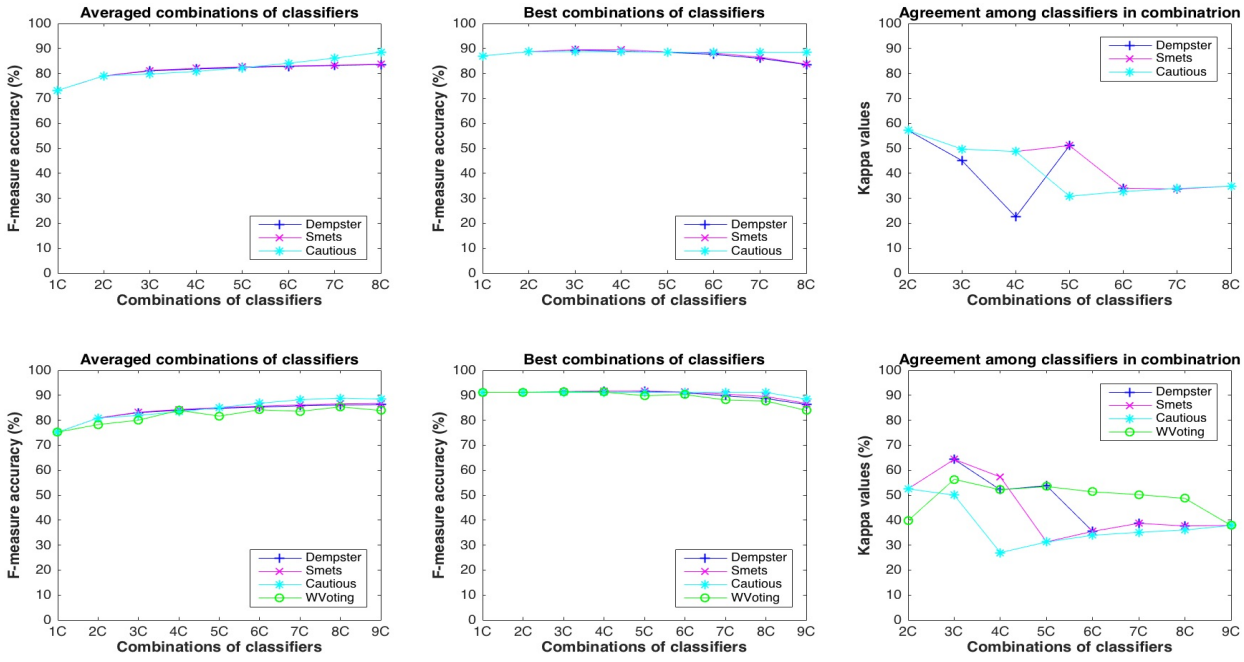

Figure 4: Accuracies of combined classifiers by four combination rules and respective Kappa values over the TripAdvisor dataset

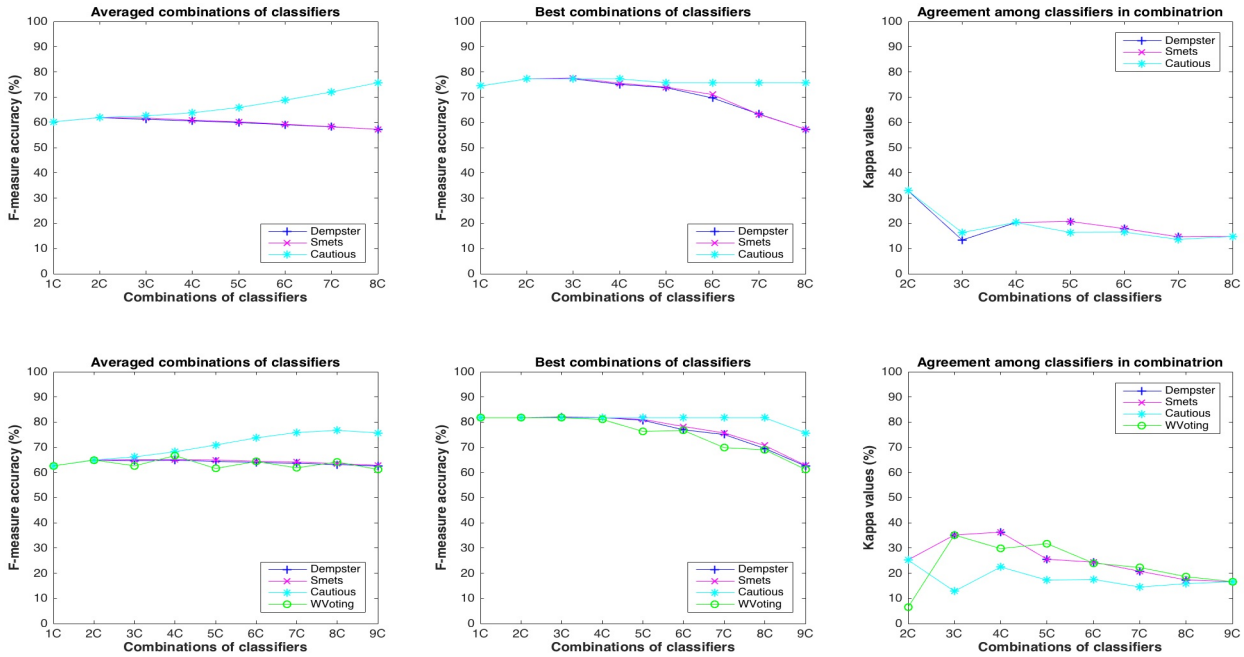

Figure 5: Accuracies of combined classifiers by four combination rules and respective Kappa values over the Yelp dataset

ensembles made from eight classifiers by Dempster's rule and Smets's rule demonstrates a decreasing trend, in particular, the combinations of classifiers by the Cautious rule start gradually decreasing from 3C until 8C at the end of combination. Meanwhile the performance of the best classifier ensembles made from nine classifiers by the Cautious rule remains relatively static until 8C. The two diagrams demonstrate the superiority of the classifier ensembles made by Dempster's and Smets's rules in the both cases. Quantitatively, the performance of the best ensembles made from eight classifiers is 0.51% better than the best individual for the three different combination rules, whereas the best combinations of nine classifiers made by Dempster's rule is 0.85% than the best individual, by Smets's rule it is 0.67% better, by the Cautious rule it is 0.15% better and by the Majority Voting it is 0.38% better. The last column in Figure 3 illustrates the κ statistic curves, representing the measurement of the agreement among the classifiers in determining the polarities of reviews. As shown in the diagram of eight classifies, the stronger agreement reach from 3C to 5C and from 6 to 8, but these curves do not show a general downward trend. One possible reason of this phenomenon could be due to the small number of words within each of reviews. For the case of nine classifiers an abnormal phenomenon occurs in the combination of two classifiers, which shows the weakest agreement − a stark contrast to the intuition and the other κ results given in Figures 1 and 2. Although the nine classifier case shows a general downward slope as the combination number of classifiers increases, the agreement between two classifiers in 2C is also weaker compared with 3C and 4C with the exception of two classifiers involved in combination by the Cautious rule, and it can be observed that the κ values decrease when combining more classifiers by the Majority Voting.

Experimental results of TripAdvisor

Figure 4 presents the experimental results of the classifier ensembles over TripAdvisor. The figures in the first column show that the averaged performance of the classifier ensembles constructed from eight and nine classifiers by the combination rules increases with small fluctuations when more classifiers being added, particularly the ensembles made by the Cautious rule appear to increase at a larger rate compared with those made by Dempster's rule and Smets's rule. Generally the both cases appear to be

compact and have a similar performance trend. Compared with the averaged performance of eight classifiers, the averaged accuracy of the classifier ensembles made by Dempster's rule is 8.72% better, by Smets's rule it is 8.89% better and by he Cautious rule it is 9.7% better. For the case of combining nine classifiers, the averaged performance of the classifier ensembles made by Dempster's rule performs 9.29% better than averaged performance of the nine individual classifiers, by Smets's rule it performs 9.6% better, by the Cautious rule it does 10.23% better, and by the Majority Voting rule 7.39% better. By contrast, the second column of Figure 4 presents the accuracies of the best classifier ensembles. The performance of the best classifier ensembles composed of eight classifier by Dempster's rule and Smets's rule increases from 1C to 3C and then decrease from 5C until 8C, while the best ensembles made by the Cautious rule just increase their accuracies from the beginning of combinations and then remain stable throughout the other combinations. For the combinations of nine classifiers, the performance of the best ensembles remains stable until 5C, but the performance of the ensemble built by the Cautious rule remain stable until 8C, and then the performance of the best ensembles starts dropping. For the case of eight classifiers, the best classifier ensemble made by Dempster's rule performs 2.36% better than the best individual, by Smets's rule, the best ensemble performs 2.45% better, and by the Cautious rule, the best ensemble does 1.63% better. In respect of nine classifiers, compared with the best individual, the performance of the best classifier ensemble constructed by Dempster 's rule increases 0.54%, by Smets's rule it increases 0.58%, by the Voting rule the performance of the best classifier ensemble increases 0.03%, and by the Cautious rule, the the difference margins with the best individual accuracy is very small.
The κ curves in Figure 4 characterise the extent of the agreement among the classifiers. In the cases of eight classifiers, the agreement among the classifiers involved in the ensembles made by Dempster's rule decreases from 2C to 4C, and increases on 5C and then decreases again. With the exception of 5C, the degree of the agreement among the classifiers making up the ensembles by Smets's rule is almost equal to that by the Cautious rule, the difference of their agreement mainly happens among three, four and five classifiers. For the nine classifier case, the agreement among two classifiers in combination is moderate strong, the strongest agreement is among three classifiers involved in constructing the classifier ensemble. However the figures show that the inconsistent agreement occurs among three, four and five classifiers for the four different combination rules, but the agreement among the classifiers involved for the Majority Voting rule is the strongest across the different groups of the classifiers on average.

Experimental results of Yelp

The experimental results over the Yelp dataset are presented in Figure 5. The averaged performance of the classifier ensembles composed from eight classifiers by Dempster's rule and Smets's rule achieves best on 2C and decreases throughout the other combinations, whereas the averaged performance of the classifier ensembles made by the Cautious rule increase until the end of combinations of eight classifiers and until 8C for the nine classifiers. The averaged accuracies of the classifier ensembles made by Dempster's rule and Smets's rule appear to be less than that of the averaged individuals for the both cases. Compared with the averaged individual, the averaged performance of the classifier ensemble made by the Cautious rule is 7.03% better. In relation to the nine classifiers, the averaged performance of the classifier ensembles made by Dempster's rule is 1.41% better than the averaged individual, by Smets's rule it is 1.78% better, and by the Cautious rule it is 8.93% better, but by the Majority Voting rule it is 0.82% better.
The second column in Figure 5 presents the best combinations of classifiers. In the case of eight classifiers that are involved in the construction of the best ensembles, the performance of the best classifier ensembles made by Dempster's rule and Smets's rules increases on 2C and 3C and then continually decrease to the end of combinations. The performance of the best ensembles made by the Cautious rule increases from 1C to 2C and remains flat until the end of combinations. For the case of nine classifiers, the accuracies of the best classifier ensembles built by Dempster's rule and Smets's rule are the same as those of the best individual from 1C to 4C, and then drop throughout the other combinations, but the accuracies of the best classifier ensembles made by the Cautious rule remain flat until 8C. Compared with the best individual, the accuracy of the best classifier ensemble made by

Dempster's rule is 2.8%, by Smets's rule it is 3.17% better and by the Cautious rule is 2.76% better. When nine classifiers participating in the construction of ensembles, the performance of the best classifier ensemble made by Dempster's rule is 0.32% better than the best individual, by Smets's rule it is 0.13% better, and the best ensemble by the Cautious rule and the Majority Voting rule almost perform the same as the best individual.
The diagrams in the last column show κ statistics, depicting the extent of the agreement among the classifiers involved in making up the best classifier ensembles. In the case of eight classifiers, the degree of the agreement three classifiers is the lowest, and then increases until 5C and then decreases towards to flat until the end of combination. For the case of nine classifiers, the agreement among the classifiers making up the best ensembles by the Cautious rule is the same as one of eight classifiers. But the degree of the agreement among the classifiers constructing the ensembles by Dempster's and Smets's rules increases until 4C and decrease throughout the other combinations, by the Majority Voting rule the situation is similar to the cases of Dempster's and Smets's rules with the exception of 2C.

Time complexity analysis on five datasets with 4 combination rules

To assess the computational complexity of the combination rules, Figure 6 presents the average time taken in constructing classifier ensembles (on left hand side) and the time spent for constructing the best classifier ensembles (on right hand side). This figure shows that the Weighted Majority Voting rule is more efficient, whereas the Cautious rule is more computational expensive among these four rules. The average time taken across the five data sets is 0.1280 second for the Majority Voting, 0.5523 for the Dempster's rule, 0.5843 for Smet's rule, and 5.6665 seconds for the Cautious rule, which are significantly small compared with the time taken for training classifiers in general. It is envisaged that the trade off between performance and efficiency is a common issue in constructing classifier ensembles and it is also extremely difficult to improve the performance of any machine learning algorithms. The preceding analysis results reveal that the proposed ensemble approach does not add much computational complexity in constructing classifier ensembles, in meanwhile improves the performance of the individual classifiers to some extent.

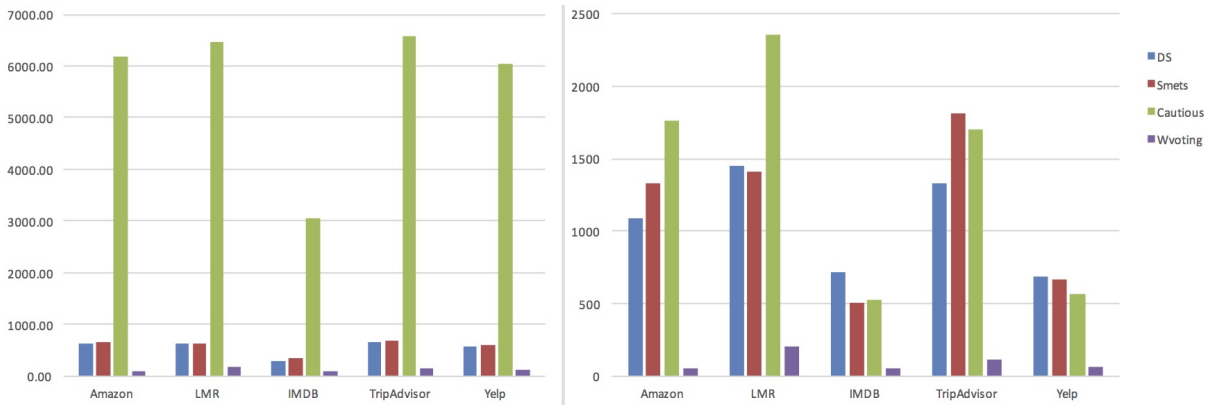

Figure 6: Time spent on combining classifiers by four combination rules: average time on the left in milliseconds and time of the best classifier ensembles on the right in milliseconds

Discussions

Summary of two group experiments

The difference between the two groups of experiments is whether SVM should be included in constructing classifier ensembles. The impact of SVM on the ensemble performance is clearly evident in

Table 4: The differences between the best individual classifiers and best classifier ensembles constructed by DS, SR, CR and WV on the five datasets in both cases of eight and nine classifiers.

Dataset	**Eight classifier**			**Nine classifier**			
	DS	SR	CR	DS	SR	CR	WV
Amazon	5.90	6.01	3.39	0.35	0.36	0.01	0.15
LMR	0.68	0.59	0.42	1.64	1.67	1.34	1.38
Rotten	0.51	0.51	0.51	0.85	0.67	0.15	0.38
Trip	2.36	2.45	1.63	0.54	0.58	0.001	0.06
Yelp	2.80	3.17	2.76	0.31	0.13	0.005	0.06
Av	2.45%	2.55%	1.74%	0.74%	0.68%	0.30%	0.46%

Tables 2 and 3. The performance of the best individuals is dominated over the datasets by NB in the case of eight classifiers, while SVM dominates the performance of nine classifiers, and the averaged accuracy with SVM is 85.61%, 4.61% better than that without SVM.

Table 4 presents the performance differences between the best individuals and the best classifier ensembles over all the datasets in the both cases presented in Tables 2 and 3, showing the improvements against the individuals. It can be seen that the best classifier ensembles outperform the best individuals to varying extent, where the accuracy improvement of the ensembles composed in the case of eight classifiers is 1.67% higher than that made from nine classifiers on average, this difference is in contrast to that the averaged best ensemble of nine classifiers is 1.43% better than that of eight classifiers (from average difference in Tables 2 and 3).

Paired t-tests have been also conducted to examine the statistical significance of the differences given in Table 4. For each of the best individuals and best classifier ensembles, we collected the classification accuracy in each fold of a data set and put these accuracies together to form a ten dimension vector of the classification accuracies, we then compute t-tests on the pairs of the vectors of the best individuals against the best classifier ensembles over the five data sets. The calculated results confirm that the paired differences for the case of the eight classifiers are statistically significant at a 95% confidence level ($p < 0.05$), however for the case of the nine classifiers, the t-test results are varied over these datasets. Specifically the t-test results confirm that the differences between the best individuals and the best classifier ensembles over the LMR dataset are statistically significant at a 95% confidence level, the differences with the best ensembles made by Dempster's and Smets's rules over TripAdvisor are statistically significant at a 95% confidence level, and the difference with the best ensemble made by Dempster's rule on Rotten-IMDB is statistically significant at a 90% confidence level. For the rest, the t-test results cannot confirm they are statistically significant at 90-94.9% confidence levels.

To understand the difference of the improvement between the two cases, we use the averaged individual classifiers as references and then compare them against the averaged ensembles and the best individuals, respectively. Table 5 presents the first comparison over the five datasets, it shows that the averaged ensembles outperform the averaged individuals to large extent, and the difference of performance between the two averages of the nine classifiers is 1.42% higher than that of the eight classifiers on average. Figure 7 illustrates the accuracies of the averaged classifiers and the best individual classifiers in the cases of eight and nine classifiers, it is interesting to observe that the difference between the best individual and the average of the nine classifiers is 1.9% higher than that of the eight classifiers on average.

These comparative analyses conjecture a possible reason of the small improvement for the combinations of the nine classifiers, that is the performance of the best individuals and averaged classifier ensembles has a big margin with reference to the performance of the averaged individual classifiers, which results from the dominant performance of SVM throughout the five datasets. This subsequently reveals a fact when the performance difference between the best classifier with the other member classifiers is in a large margin, the classifier ensemble made by combining some of the others would achieve a smaller improvement of performance, which is consist with the general perception in the literature (Bi, 2012).

Table 5: The differences between the averaged individual classifiers and averaged classifier ensembles constructed by by DS, SR, CR and WV on the five datasets in both cases of eight and nine classifiers.

Dataset	**Eight classifier**			**Nine classifier**			
	DS	SR	CR	DS	SR	CR	WV
Amazon	3.03	2.91	7.20	5.24	5.60	8.81	3.79
LMR	7.67	7.21	7.78	9.03	8.84	9.27	8.81
Rotten	10.61	9.09	2.55	10.35	9.83	4.77	4.86
Trip	8.72	8.89	9.70	9.29	9.60	10.23	7.39
Yelp	-0.49	-0.30	7.03	1.41	1.78	8.93	0.82
Av	5.91%	5.56%	6.85%	7.06%	7.13%	8.40%	5.13%

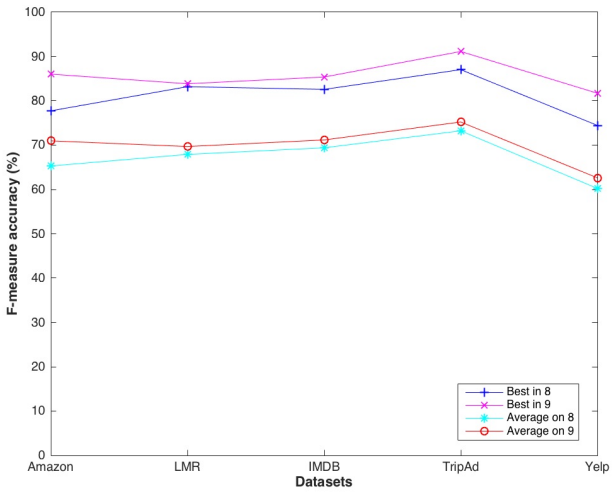

Figure 7: Performance of the best and average individual classifiers over the five datasets

Table 6: Correlation between the best ensemble accuracies and κ statistics of the component classifiers making up the best ensembles constructed by DS, SR, CR and WV on the five datasets.

Dataset	**Eight classifier**			**Nine classifier**			
	DS	SR	CR	DS	SR	CR	WV
Amazon	0.3487	**0.7129**	**0.9309**	**0.6757**	**0.7041**	0.1383	0.2996
LMR	**0.5847**	0.5483	**0.8807**	0.4892	0.4393	0.1585	**0.6417**
Rotten	-0.1326	-0.1975	-0.5533	0.3412	0.5270	0.1309	0.3313
Trip	0.3125	**0.6798**	**0.9595**	**0.6389**	0.3409	0.0059	**0.5845**
Yelp	0.5078	**0.5871**	**0.6378**	**0.8155**	**0.7913**	0.1146	0.2979
Av	0.3242	0.4661	**0.5711**	**0.5921**	**0.5605**	0.1096	0.4310

Correlation between performance and κ statistics

The last columns in Figures 1 - 5 illustrate the κ statistics calculated from the groups of classifiers which make up the best respective classifier ensembles, they quantifiably indicate how strong the member classifiers of an ensemble agree each other in determining sentiment polarities of reviews. For each of the best ensembles constructed by Dempster's rule, Smets's rule and the Majority Voting, the κ statistics among the member classifiers is decreasing in general with a variation in two classifiers (2C), whereas by the Cautious rule, the κ values vary within a small range, particularly after 4-5 classifiers. This result intuitively supports that the agreement among the smaller number of classifiers is approximately higher than that of the larger number of classifiers in combination. This also reveals a fact that the lower κ statistic agreement has the less effect on constructing the best classifier ensembles, as shown in the diagrams that the κ statistic values in the case of eight classifiers are smaller than those of nine classifiers. To quantify this effect, a correlation between the accuracies of the best classifier ensembles and the degrees of agreement among the respective classifiers are computed by the Spearmen's rank method and the resulting coefficients are presented in Table 6. The table shows the performance of the ensembles made by the Cautious rule is more correlated to the agreement among the respective classifiers in the case of eight combinations, the correlation between the ensemble accuracies made by Dempster's rule and the κ statistics of component classifiers is strong over Amazon, TripAdvisor and Yelp in the case of nine classifiers, and the performance of the ensembles constructed by Smets's rule is more correlated with the agreement of the component classifiers in the both cases. However the number of strong correlation coefficients are not sufficient to conclude that the ensemble performance is closely correlated with the agreement of component classifiers in classifying instances in the form of κ statistics.

For the case of eight classifiers, it is also worth noticing that there are the negative correlation values of DS, SR and CR rules in the Rotten-IMDB dataset in Table 6. Examining the performance of the best classifier ensembles by the three combination rules along with the κ statistic curves as shown in Figure 3, the accuracy of the best classifier ensembles decreases after 2C, whereas the the κ statistic curves increase. This inverse relationship could result in these negative correlation values, which might be caused by a small number of words within each of reviews.

Conceptual comparison

Many machine learning methods have been studied for sentiment classification in the literature. The most popular sentiment classification study is one conducted on the document-level by Pang, et al. (2002). The authors compared NB, ME, and SVM three leaning algorithms and concluded that SVM generally achieved the best results. Since then there are many following up studies attempting to improve the performance of machine learning-based sentiment classification. Table 7 presents two lists of machine learning techniques that have been reviewed by Injadat et al., (2016), and Ravi and Ravi (2015), respectively, these two lists have a more than 50% overlap. It can be seen that six machine learning algorithms used in our study have not been listed in these two survey papers, whereby widening the scope of applications of learning algorithms for sentiment classification.

Although a large number of experimental results have been reported in publications, those evaluation

Table 7: Machine learning techniques used in sentiment classification

No	**Review 1** (Ravi and Ravi, 2015)	**Review 2** (Injadat et al., 2016)
1	AdaBoost	Boosting
2	Artificial Neural Network (ANN)	RBF NN, NN
3	Bayesian Networks (BN)	NB
4	Decision Trees (DT)	DT
5	Fuzzy	Fuzzy logic
6	K-Means	K-Medoids
7	Linear-Regression	Linear Regression
8	Logistic Regression	Logistic Regression
9	Maximum Entropy	Maximum Entropy (ME)
10	Support Vector Machine	SVM, SVM-SMO
11	Linear Discriminant Analysis (LDA)	LDA
12	Density Based Algorithm	Dictionary Based Approaches
13	Wrapper	Random forest
14	Apriori	SVR
15	Genetic Algorithm (GA)	CRF and rCRP
16	Hierarchical Clustering (HC)	Ontology
17	Markov	Rule miner
18	Novel	EM
19	k-nearest Neighbors (kNN)	

work have been conducted with different experiment settings, different versions of data sets and even with different performance measurements. As a result, a direct comparison with those previous studies is infeasible. Taking Table 7 into account, Table 8 presents thirteen typical studies attributed with *Method*, *Feature*, *Corpus* and *Accuracy*, it provides a conceptual comparison on the performance of state of the art machine learning methods used for sentiment classification to date, where the highest accuracy is 93.2% obtained by Yadav et al. (2019). However the highest accuracy obtained on IMDB by a 2-partition validation is in stark contrast with the classification accuracy of 50.9% on Yelp in the same study (Baly et al., 2016), which makes the generalisation of the Meta-Framework developed questionable.

Of particular relevant to our study is the ensemble of feature sets and classification algorithms conducted by Xia et al., (2011). The authors used the three types of ensemble methods: fixed combination rules, weighted combination and meta-classifier trained combination to construct classifier ensembles and conducted the document-level experiments with a 5-fold cross validation, the dataset consists of 1000 positive and 1000 negative processed reviews derived from the Cornell movie-review corpora (Pang and Lee, 2004), as well as product reviews taken from Amazon.com about four product types of *book*, *DVD*, *electronics* and *kitchen* used by Jo and Oh (2011), each of which comprises 1000 positive and 1000 negative reviews. They used the SVM with unigrams and the SVM jointed by word relation features as two baselines, and then achieved 83.12% accuracy compared with 79.99% on Baseline 1 with a 3.13% improvement, and 85.58% accuracy on average over Baseline 2 by a 3.29% improvement.

Another relevant study is published by Wang et al. (2014), the authors compared the performance of 20 combined classifiers. These algorithms are taken from the Weka 3.7.0. They experimented on ten data sets with a ten-fold cross validation, the datasets include the movie dataset comprising 1000 positive and 1000 negative reviews (Pang and Lee, 2004), and nine datasets taken from the study (Whitehead and Yaeger, 2009). The empirical results show that the best accuracies of the ensemble methods on six datasets are better than the best individuals, including the best accuracy 85.48% on the Camp dataset, 85.97% on the *doctor* dataset, 70.26% on the *drug* dataset, 92.62% on the *laptop* dataset, 82.90% on the *movie* dataset and 72.13% on the *music* dataset. The best averaged accuracy of these best classifier ensembles is 81.5% compared with the averaged best individuals 79.6%, which is a 1.9% increase on these datasets.

The size of the datasets used for our study is generally larger than those discussed above and in Table 8, and these datasets are across the domains of movies, products, travelling, hotels and

Table 8: Conceptual comparison with existing studies

Method	**Feature**	**Corpus**	**Accuracy**
Supervised (Pang et al., 2002)	Bigrams, unigrams	IMDB	82.9%
Cut-based classification (Pang and Lee, 2004)	Graphs	Rotten-IMDB	87.15%
Unsupervised (Turney, 2002)	Predefined phrase patterns, POS	Product, movie and travel reviews	74.4%
Entropy weighted genetic algorithm (Varathan et al., 2017)	Stylistic, syntactic features	Movie reviews	91.70%
Gated recurrent neural Networks (RNN) (Baly et al., 2016)	n-word, n-character	IMDB	92.1%
CNN +RNN (Yadav et al. 2019)	Word vector	IMDB	93.2%
Unsupervised and supervised (Maas et al., 2011)	Word vectors	Very large movie reviews	88.89%
Extended latent dirichlet allocation (Jo and Oh, 2011)	Unigrams	Amazon and Yelp	84.00-86.00%
Hybrid classifiers (Prabowo and Thelwall, 2009)	Unigrams	Product and movie reviews, and MySpace comments	90.45%
Class sequential rules (Tan et al., 2012)	Dependency bigrams	IMDB	84.97%
Lexicon-enhanced method (Dang et al., 2010)	n-grams and syntactic features	Product reviews	84.21%
CNN, Long Short Term Memory (LSTM) (Wang et al. 2018)	Word vectors	Amazon	84.7-87.9%
Gated Recurrent Unit (GRU), RNN (Li, et al. 2019)	Word Embedding	Yelp	67.30-74.12%

restaurants, which are more comprehensive. Regardless of how the performance of those methods were measured, in comparison with our two cases of eight and nine classifiers, the first performance is 83.55% with a 2.55% improvement on average, and the second result is 86.35%, 0.74% better than the best individual on average. Therefore we would say that our results are competitive to those results reported in the literature and involve relatively simple process of features.

Experimental comparison

In order to reinforce our studying results above, we have carried out comparative studies against the AdaBoosting ensemble method based on the algorithms of NB, ME, DT, kNN and SVM used in the study (Wang et al.,2014) and a latest deep learning algorithm of the Transformer (Nandan, 2020) over the five datasets. All the algorithms boosted are drawn from the Weka package in Java with the default settings as done previously, but the Transformer is taken from the most used deep learning framework Keras in Python. The Transformer has a set of hyperparameters. For this comparative study, we select six important hyperparameters with initial values, including *embedding dimension* = {32, 64, 128, 256}, *attention heads* = {1, 2, 3, 4}, *hidden layer size* = {32, 64, 128, 256}, *batch* = {32, 64, 128, 256}, *epochs* = {10, 20, 50} and *number of words in each batch* = {100, 200, 300, 400}, we then tuned these values on one dataset. The set of tuned optimal values, {32, 1, 32, 128, 20, 300}, have been used in the experiments.

Table 9 presents the accuracies of the base classifiers and classifier ensembles boosted over the five data sets. From the left column of the table, it can be observed that on average TNN performs best and SVM performs second best. The right column illustrates that NB performs best, DT performs second best, and the improvement of the boosting accuracies are 4.59% and 5.88% for NB and DT, respectively, on average. The experimental results demonstrate that the boosting performance does not rely on that of the base classifiers, which is somehow consistent with the experimental results reported by Wang, et al.(2014). On the other hand, comparing the results with those in Table 3, the performance of the best classifier ensemble made by Dempster's rule is 2.02% better than that of the best boosted classifier of NB on average, but it could outperform the Transformer.

Comparing the performance of the Transformer with the best classifier ensembles further, we found that the Transformer performs better than the best classifier ensembles with 1.17% on average, specifically 0.12% better on *Large-Movie-Review*, 2.48% on *Rotten-IMDB*, 0.89% on *TripAdvisor* and 2.48% on *Yelp*, respectively, but not on *Amazon*. It is not surprised that the Transformer achieved a 2.48% better accuracy on the Yelp as this dataset was used for tuning the hyperparameters. Apart from this, the performance on *Rotten-IMDB* could conjecture that the Transformer preferred the shorter length of reviews. It should be also reiterated that such better performance was achieved through tuning a set of hyperparameters. When the parameter values changed, the performance of the Transformer could be degraded. For example, the folder number was changed to 5, the Transformer performed 0.87% better than the best classifier ensembles on average.

For comparing the computational complexity of the AdaBoosting and the combination rules, we have estimated the time spent for training the NB, ME, kNN, DT and SVM base classifiers and boosting them by AdaBoostingM1 separately, and took the differences between two computation times and then averaged them over the five data sets. The left column of Table 10 shows the average time spent for constructing classifier ensembles by the combination rules and AdaBoosting, and the right column exhibits the time spent in making up the best classifier ensembles. As seen that the average time spent by Dempster's and Smets's rules and Majority Voting is less than one second in constructing classifier ensembles, whereas the AdaBoostingM1 takes 10285.91 seconds on average, which is 1484.26 times more than the combination rules. Additionally the running time of the combination rules is 0.92 second in constructing the best classifier ensembles on average, which is 3231.84 times faster as AdaBoostingM1 does.

These comparative studies confirm the competitiveness of our proposed ensemble method either in the classification performance or running time, they also support an assertion that utilising the evidential combination rules would be able to harness the strength of the multiple learning algorithms, which outperforms the AdaBoosting approach that derives the strength from a single algorithm through self adapting over the various distributions of training data.

Table 9: The accuracies of the base individual classifier generated by NB, ME, DT, kNN, SVM and TNN and the accuracies of the boosted classifiers on the respective NB, ME, DT, kNN and SVM by AdaBoostM1 (AB) over the five datasets .

Dataset	**Individual classifier**						**AB**				
	NB	ME	DT	kNN	SVM	TNN	NB	ME	DT	kNN	SVM
Amazon	77.75	79.55	76.05	56.55	86.02	86.20	**83.54**	79.55	**83.73**	56.55	85.62
LMR	83.20	80.05	73.80	59.35	83.85	85.70	**84.78**	80.03	**80.29**	59.34	83.21
Rotten	76.20	82.85	77.50	61.60	85.37	88.70	**84.08**	79.29	**81.09**	**63.95**	83.49
Trip	87.05	86.55	81.55	69.75	91.13	92.60	**90.75**	86.56	**88.22**	69.74	90.66
Yelp	74.45	76.60	70.65	51.80	81.71	84.50	**78.46**	76.60	**75.62**	51.77	81.02
Av	79.73%	81.12%	75.91%	59.81%	85.61%	87.54%	84.32%	80.41%	81.79%	60.27%	84.80%

Table 10: The average time taken in building classifier ensembles and time taken in building the best classifier ensembles in seconds on the five datasets.

Dataset	**Time taken for average**					**Time taken for best**				
	DS	SR	CR	WV	AB	DS	SR	CR	WV	AB
Amazon	0.63	0.66	6.18	0.09	9116.15	1.09	1.32	1.75	0.05	2835.66
LMR	0.61	0.64	6.45	0.17	14269.92	1.44	1.41	2.35	0.19	3999.36
Rotten	0.31	0.34	3.06	0.08	3958.88	0.71	0.50	0.52	0.05	1525.47
Trip	0.65	0.68	6.59	0.16	14646.41	1.33	1.81	1.70	0.11	4632.02
Yelp	0.56	0.59	6.03	0.13	9438.18	0.68	0.67	0.56	0.06	1873.93
Av	0.55	0.58	5.67	0.13	10285.91	1.05	1.14	1.38	0.10	2973.29

Conclusions

In this paper we propose to adapt the triplet belief functions to represent sentiment classification outputs and present the development of the formulas for aggregating multiple triplet functions and simple support functions for constructing classifier ensembles with the three evidential combination rules. As seen the construction of the classifier ensembles here is not self adapted through manipulating the data and variables on which they are trained as the AdaBoosting method does, rather retaining the individual strength of classifiers through training and complementing one another in combination. The work also provides a systematic study on the utilization of conventional text categorization methodology in evaluating the proposed method for sentiment classification, and the comparative analyses and studies with the baseline combination rule of Majority Voting, AdaBoosting and state of the art deep learning algorithm of the Transformer, the empirical results demonstrate our method is competitive in the application of sentiment classification.

Examining the compositions of the best classifier ensembles, it is found that the best classifier ensembles built by the evidential combination rules comprise 2-5 base classifiers, this means fact that combining more classifiers represented by the triplet mass functions into an ensemble classifier might not result in an improvement of individual performance. The reason for that could be the combination of more classifiers would not effectively make the best ensemble to arrive at a convergent consensus as witnessed in the κ statistics in the diagrams. This finding is consistent to the observation obtained by Bi et al., (2008). With a further examination, it is also noticed that the best base classifier is always the constituent in the best classifier ensembles. As opposed to this, the second best constituent classifier plays in a varied role in constructing the best classifier ensembles, what roles they play are depending the combination rules and the datasets. Only in the case of nine classifiers, the best ensembles made by Dempster's and Smets's rules comprise the best and second best base classifiers across the five datasets. According to our previous studies (Burns, et al. 2011), the performance of sentiment classification on unbalanced datasets was better than that on balanced datasets. The empirical results of this study demonstrate the various results somehow, which are reflected on the results of all the classifier ensembles constructed by 4 combination rules, AdaBootM1 and the Transformer. The performance on the datasets of Large-Movie-Review and Rotten-IMDB could provide some insights into the role of the distribution of balanced classes. It might be impossible to draw a conclusion of considerable impact on

the performance of classifier ensembles.
The classification task of sentiment polarities actually involves three possible outcomes: positive, negative and neutral. To avoid additional training for the neutral category, we formulate the sentiment classification as a dichotomous classification and treat the neutral category as the form of a frame of discernment composed of positive and negative. Although a frame might absorb a large probability after aggregating triplet belief functions, it would be ignored in determining positive or negative category as the neutral can be regarded as a 'virtual' boundary − a new and innovative mechanism − to separate positive from negative category in which we are only interested.
The method proposed in this work, to our knowledge, is the first attempt application in sentiment classification. Here a classifier ensemble would represent a collection of 'experts', with varying expertise and specialization on characterizing sentiment words on which they have been trained and each constituent classifier has own belief degrees (in the form of probabilities) on deciding categories for review instances. The classification decision on the instances is made through harnessing the strengths of the classifier ensembles, hence making the decision more robust in addition to improving overall classification accuracy to some extent. Therefore this study fills the gap of using evidential combination rules to aggregate binary classifier outputs for sentiment classification, providing a new approach to employing classifier ensembles for sentiment classification tasks under a data-driven framework. A further study on combining the latest deep-learning models using our framework will be reported in a future publication.

References

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Commun. ACM*, 56(4):82–89.

Javier Rodríguez-Vidal, Julio Gonzalo, J., Plaza, L. & Anaya-Sánchez, H.(2019). Automatic detection of influencers in social networks: Authority versus domain signals. *JASIST*, 70(7):675–684.

Varathan, K., Giachanou, A. & Crestani, F. (2017). Comparative opinion mining: A review. *JASIST*, 68(4):811–829.

Pang, B., Lee, L. & Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, (pp 79–86).

Bi, Y. (2014). Evidential fusion for sentiment polarity classification. In *Belief Functions: Theory and Applications - Third International Conference, BELIEF 2014, Oxford, UK, September 26-28, 2014*, (pp 365–373).

Bi, Y., Mulvenna, M. D., & Jurek, A. (2014). Sentiment classification by combining triplet belief functions. In *Knowledge Science, Engineering and Management - 7th International Conference, KSEM 2014, Sibiu, Romania, October 16-18, 2014*, (pp 234–245).

Burns, N., Bi, Y., Wang, H. & Anderson, T. J. (2011). Sentiment analysis of customer reviews: Balanced versus unbalanced datasets. In *Knowledge-Based and Intelligent Information and Engineering Systems - 15th International Conference, KES 2011, Kaiserslautern, Germany, September 12-14, 2011, Proceedings, Part I*, (pp 161–170).

Burns, N. Bi, Y., Wang, H. & Anderson, T. J. (2019). Enhanced twofold-lda model for aspect discovery and sentiment classification. *IJKBO*, 9(4):1–20.

Bi, Y., Guan, J., & Bell, D. A. (2008). The combination of multiple classifiers using an evidential reasoning approach. *Artif. Intell.*, 172(15):1731–1751.

Bi, Y. (2008). An efficient triplet-based algorithm for evidential reasoning. *Int. J. Intell. Syst.*, 23(4):483–516.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton.

Denoeux, T. (2008). Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence. *Artif. Intell.*, 172(2-3):234–264.

Injadat, D., Salo, F., & Nassif, A., B. (2016). Data mining techniques in social media. *Neurocomput.*, 214(C):654–670.

Ravi K. & Ravi, V. (2015). A survey on opinion mining and sentiment analysis. *Know.-Based Syst.*, 89(C):14–46.

Nandan, A. (2020). Text classification with Transformer. https://keras.io/examples/nlp/text_classification_with_transformer/

Turney, P. D. (2002). Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, (pp 417–424).

Sebastiani, F.(2002) Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47.

Pang, B. & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

McDonald, R., Hannan, K., Neylon, T., Wells M. & Reynar, J. (2007). Structured models for fine-to-coarse sentiment analysis. In *Proceedings of ACL-07, the 45th Annual Meeting of the Association of Computational Linguistics.*

Jijkoun, V., de Rijke, M. & Weerkamp, W. (2010). Generating focused topic-specific sentiment lexicons. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, Stroudsburg, PA, USA. Association for Computational Linguistics, (pp585–594)

Xia, R., Zong, C. & Li, S. (2011). Ensemble of feature sets and classification algorithms for sentiment classification. *Inf. Sci.*, 181(6):1138–1152.

Apté, C., Damerau, F. & Weiss, S. M. (1994). Automated learning of decision rules for text categorization. *ACM Trans. Inf. Syst.*, 12(3):233–251.

Lewis. D. D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '92, New York, NY, USA, (pp 37–50).

Schütze, H., Hull, D. A. & Pedersen, J. O. (1995). A comparison of classifiers and document representations for the routing problem. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, New York, NY, USA (pp 229–237).

Lane, P., Clarke, D., & Hender, P. (2012). On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data. *Decis. Support Syst.*, 53(4):712–718.

Wang, G. Sun, J., Ma, J., Xu, K, & Gu, J. (2014). Sentiment classification: The contribution of ensemble learning. *Decision Support Systems*, 57:77–93.

Tan, L., Na, J., Theng, Y. L. & Chang, K. (2012). Phrase-level sentiment polarity classification using rule-based typed dependencies and additional complex phrases consideration. *J. Comput. Sci. Technol.*, 27(3):650–666.

Whitehead M. & Yaeger, L. S. (2008). Sentiment mining using ensemble classification models. In *Innovations and Advances in Computer Sciences and Engineering, Volume I of the proceedings of the 2008 International Conference on Systems, Computing Sciences and Software Engineering (SCSS), part of the International Joint Conferences on Computer, Information, and Systems Sciences, and Engineering, CISSE 2008, Bridgeport, Connecticut, USA*, (pp 509–514).

Smets, P. & Kennes, R.(1994). The transferable belief model. *Artif. Intell.*, 66(2):191–234.

Witten, I. H., Eibe, F., & Hall, M. A. (2011). *Data mining: practical machine learning tools and techniques, 3rd Edition*. Morgan Kaufmann, Elsevier.

Jo, Y. & Oh, A. H. (2011). Aspect and sentiment unification model for online review analysis. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, New York, NY, USA (pp 815–824).

IMDB. (2006). Internet movie database (imdb) archive from https://www.archiveteam.org.

Tripadvisor achieve (2006). https://www.tripadvisor.co.uk

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y. & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, Stroudsburg, PA, USA. Association for Computational Linguistics, (pp142–150)

Zhitomirsky-Geffet, M., Bar-Ilan, J. & Levene, M. Categorical relevance judgment. (2018). *JASIST*, 69(9):1084–1094.

Vilares, D., Alonso, M., A. & Gómez-Rodríguez, C. (2015) On the usefulness of lexical and syntactic processing in polarity classification of twitter messages. *JASIST*, 66(9):1799–1816.

Baly, R., Hobeica, R., Hajj, H., El-Hajj, W., Shaban, K. B. & Al-Sallab, A. (2016). A meta-framework for modeling the human reading process in sentiment analysis. *ACM Trans. Inf. Syst.*, 35(1):7:1–7:21.

Prabowo, R. & Thelwall, M. (2009). Sentiment analysis: A combined approach. *J. Informetrics*, 3(2):143–157.

Dang, Y., Zhang, Y., & Chen, H (2010). A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, 25(4):46–53.

Bi, Y.(2012) The impact of diversity on the accuracy of evidential classifier ensembles. *Int. J. Approx. Reasoning*, 53(4):584–607.

Pang, B., & Lee, L. (2005) Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *CoRR*, abs/cs/0506075.

Prabowo, R. & Thelwall, M. (2009) Sentiment analysis: A combined approach. *J. Informetrics*, 3(2):143–157.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A., & Potts, C. (2013) Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*.

Gonçalves, P., Dalip, D.H., Costa, H., Gonçalves, M.A., and Benevenuto, F. (2016). On the combination of "Off-The-Shelf" sentiment analysis methods. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing (SAC'16)*, 1158–1165. Pisa, Italy.

Philipe F. M, Daniel H. D., Manoel M. J., Marcos A. G., Fabrício B. (2019). 10SENT: A Stable Sentiment Analysis Method Based on the Combination of Off-The-Shelf Approaches. *Journal of the Association for Information Science and Technology*, 70(3):242–255.

Yogarajan, V., Gouk, H, S., T, M. (2020). Comparing High Dimensional Word Embeddings Trained on Medical Text to Bag-of-Words For Predicting Medical Codes. *In NT Nguyen, K Jearanaitanakij, A Selamat & B Trawiski (eds), Intelligent Information and Database Systems : ACIIDS 2020*. Lecture Notes in Computer Science, vol. 12033, Springer, Cham, pp. 97-108.

Wadawadagi, R. & Pagi, V. (2020). Sentiment analysis with deep neural networks: comparative study and performance assessment. *Artificial Intelligence Review*, in press.

Hemmatian, F. & Sohrabi, M. K. (2019). A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review*, 52, 495–1545

Stuart J. Blair, Yaxin Bi, Maurice D. Mulvenna (2020) Aggregated topic models for increasing social media topic coherence. Applied Intelligence. 50(1): 138-156 (2020)

Wei Zhao, Ziyu Guan, Long Chen, Xiaofei He, Deng Cai, Beidou Wang, and Quan Wang (2018). Weakly-Supervised Deep Embedding for Product Review Sentiment Analysis. *IEEE Transactions on Knowledge and Data Engineering*, 30 (1), 185-197.

Bo Li, Zehua Cheng, Zhenghua Xu, Wei Ye, Thomas Lukasiewicz, Shikun Zhang (2019). Long text analysis using sliced recurrent neural networks with breaking point information enrichment. *Proceedings of the 2019 IEEE international conference on acoustics, speech and signal processing*, ICASSP 2019, Brighton, UK, vol 124, pp 51-60

Tien Thanh Nguyen, Mai Phuong Nguyen, Xuan Cuong Pham, Alan Wee-Chung Liew (2018). Heterogeneous classifier ensemble with fuzzy rule-based meta learner. *Information Sciences*, 422, 144-160